(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: March 14, 2000

Application Number: Japanese Patent Application
No. 2000-070915

Applicant(s): HITACHI SOFTWARE ENGINEERING CO., LTD.

January 26, 2001

Commissioner,
Patent Office Kozo OIKAWA (seal)

Certificate No. 2001-3002014

TITLE OF THE INVENTION

METHOD FOR DISPLAYING RESULTS OF HYBRIDIZATION EXPERIMENT

BACKGROUND OF THE INVENTION

1. FIELD OF THE INVENTION

The present invention relates to a method for displaying results of hybridization experiments in which a biochip is used to hybridize a sample biopolymer with a probe biopolymer with a known sequence.

2. DETAILED DESCRIPTION OF THE PRIOR ART

Biochips, also known as DNA micro arrays, have been developed to simultaneously quantify various biopolymer species, such as DNA sequences, that are present in a sample in different volumes. The technology is overviewed in Vivian G. Cheung *et al.*, "Making and reading microarrays," *Nature Genetics Supplement*, vol.21, January 1999.

In a typical biochip technique, different probe biopolymers, for example, DNA molecules, are immobilized on a surface of a support such as glass slides and, through hybridization, selectively bind to different labeled biopolymers, for example, DNA sequences, in a sample. Specific sample biopolymers can be quantified based on the amounts of markers that have been selectively coupled to the probe biopolymers via sample biopolymers hybridized to the probe biopolymers. This principle makes it possible to quantify many different sample

biopolymers at a time by immobilizing many different probe biopolymers on the same support.

In order for two DNA sequences to hybridize, the two sequences need to have base sequences complementary, or nearly complementary, to one another. When hybridized, the complementary strands have a high binding energy and are stable at a certain temperature. This binding energy varies depending on the length and base composition (GC content) of the strand. Two hybridized strands with partially non-complementary sequences can also have a sufficiently high binding energy when they contain complementary regions of sufficient lengths. This means that there is a chance that sample DNA molecules of the same type bind to two different types of DNA probes that are very similar to one another. It is known that the likelihood that this unintended hybridization (miss-hybridization) occurs varies depending on the conditions of hybridization experiments.

A type of biochips that uses synthetic short DNA strands as DNA probes is known (oligonucleotide array). In this type of biochips, DNA molecules with sequences similar to respective subject DNA probes are synthesized and used to serve as DNA probes for comparison so as to determine if the hybridized sample DNA is the intended sequence as the target of the subject DNA probe. This technique is reviewed by Robert J. Lipshutz *et al* (Robert J. Lipshutz *et al.*: High density synthetic

oligonucleotide arrays, Nature Genetics Supplement, Vol.21, January 1999).

However, in biochips that use longer DNA molecules, such as cDNA, as a probe biopolymer, no effective technique is known that can evaluate the results of hybridization using DNA sequence data.

The Smith-Waterman method is a known technique for searching for regions with highest homology between two different DNA sequences (Smith, T. F. and Waterman, M.S.: J. Mol. Biol. 147, 195-197, 1981). Also, methods are known such as BLAST that allow for a fast search for a target sequence having a high homology with a DNA sequence of interest (key sequence) among many different DNA sequences (targets) (Altschul *et al.*, Nucleic Acids Res., 25, 3389-3402, 1997). Many other algorithms have been developed for the same purpose. In these approaches, the degrees of homology between two DNA sequences are expressed by indices such as "homology score," which is based on the scores used in the search for high homology regions between the two DNA sequences, or by "matching rate," which is based on the proportion of the complementary DNA portions in the region (these indices, each representing the degree of homology, are collectively referred to as "similarity score," hereinafter.).

In a technique widely used for data analysis of biochip experiments, subject DNA proves are statistically classified

(i.e., clustered) based on the changes in levels of hybridization in a plurality of biochips. In the expression analysis for yeast conducted by P. Brown's group of the Stanford University, a DNA sample was prepared at each stage of cell development in a time-sequential manner and the samples were each hybridized to separate biochips. Types and the amounts of DNA sequences present in the DNA samples were determined for each stage. The DNA sequences (DNA probes) were then clustered based on the changes in amounts at each stage (Michel B. Eisen *et al.*: Cluster analysis and display of genome-wide expression patterns: Proc. Natl. Acad. Sci. (1998) Dec 8, 95 (25), 14863-8). The results are displayed in a tree diagram obtained from the clustering that indicates the order of clusters in the DNA sequence and the distances between the clusters. The results also include information about the DNA sequences (e.g., name, definitions, or the like) and hybridization patterns indicating the levels of hybridization for each DNA sequence on each of the biochips.

At present, no practical approach is known for determining if a probe biopolymer has been accurately hybridized to a sample biopolymer of interest, and accordingly, there is a need for such a method.

SUMMARY OF THE INVENTION

The present invention is devised to satisfy such a need.

Accordingly, it is an object of the present invention to provide a method for displaying information concerning the accuracy of hybridization experiments using biochips in a manner that is visually easy to understand.

In one embodiment of the present invention, a similarity score is calculated from the base sequences of subject probe biopolymers. The calculated score is represented by, for example, square patterns (*i.e.*, similarity patterns) having varying depths in a color. The similarity pattern, probe biopolymer data and the hybridization-level data are displayed side by side so that they can be compared with each other. The comparison makes it possible to visually confirm whether a probe biopolymer with a similar base sequence to that of an object probe biopolymer has a similar hybridization level. As a result, it can be easily known whether an unexpected hybridization reaction has taken place. Also, by simultaneously displaying the hybridization-level information of the subject probe biopolymers for multiple biochips, it is possible to determine if there is any biochip with improper hybridization. The similarity pattern can be presented in a matrix-like form by arranging the subject probe biopolymers vertically and horizontally (*i.e.*, similarity pattern matrix). This makes the displayed image more intuitive.

Further, the results of cluster analyses performed on multiple biochips and the similarity pattern matrix may be

arranged side by side to make it possible to determine whether the clusters have been separated based on the biological properties, rather than physical properties, of the base sequences.

Accordingly, the present invention provides a method for displaying the results of a hybridization experiment in which a plurality of probe biopolymers immobilized on a biochip are hybridized to a sample biopolymer. The method is characterized in that the information obtained in the hybridization experiment about the hybridization level on each of the probe biopolymers is displayed together with a similarity score representing the similarity of base sequences between each of the probe biopolymers.

Also, different depths of a color may be assigned to different values of the similarity score for the purpose of displaying. Alternatively, different depths of a color may be assigned to different values of the similarity score, and subject probe biopolymers are further arranged horizontally and vertically to form a matrix for the purpose of displaying.

The information about the hybridization level may be displayed by assigning different depths of a color to different values of the hybridization level, or by providing spot images of respective probe biopolymers.

In an effective displaying method in accordance with the present invention, probe biopolymer data (e.g., name,

definitions, or the like), hybridization levels and similarity scores are displayed side by side by sorting them based on the values of the similarity score between specific one of the probe biopolymers and each of the probe biopolymers. It is also effective to display the hybridization levels obtained from a plurality of biochips side by side.

Further, the profile of the changes in the hybridization levels of the subject probe biopolymers on said plurality of biochips may be statistically analyzed (e.g. by using cluster analyses), and the results of the analysis are displayed together with the results of clustering the probe biopolymers side by side.

The similarity scores are calculated from the base sequence information of the subject probe biopolymers, and square patterns each provided with a different depth in a color are assigned to the calculated similarity scores. The similarity scores are calculated for all of the possible combinations of the subject probe biopolymers, and corresponding squares are arranged in a matrix-like manner to serve as a similarity score matrix, which is displayed with the hybridization-level information for each of the subject probes so that the matrix and the hybridization information are arranged side by side. The matrix and the hybridization-level information are displayed in the order given by the sorting by the similarity scores between an object probe biopolymer and

each of the other probe biopolymers. As a result, the matrix and the hybridization-level information are sorted in the order of decreasing similarity with respect to the object probe. Thus, it is possible to determine if unintended hybridization has occurred by observing the hybridization-level information in the proximity of the objective probe. Also, by selecting the information to be displayed with the similarity score matrix, the verification of the accuracy of the hybridization is possible in wider ranges.

BRIEF DESCRIPTION OF THE DRAWINGS

These as well as other features of the present invention will become more apparent upon reference to the drawings in which:

Fig.1 is an illustration schematically showing an arrangement of a system in accordance with the present invention;

Fig.2 is a flow chart schematically showing a flow of processes in one embodiment of the displayed image of the sequence similarity patterns in accordance with the present invention;

Fig.3 is a diagram showing an example of a data structure stored in a biochip data storage unit;

Fig.4 is a diagram showing an example of biochip data stored in a biochip table;

Fig.5 is a diagram showing an example of a DNA probe data stored in a DNA probe table;

Fig.6 is a diagram showing an example of hybridization-level data stored in a hybridization-level table;

Fig.7 is a diagram showing an example of calculated similarity scores;

Fig.8 is a descriptive illustration of a similarity pattern;

Fig.9 is a diagram showing an example of how a similarity pattern is generated;

Fig.10 is a descriptive illustration of a similarity pattern matrix;

Fig.11 is a diagram showing an example of how a hybridization-level pattern is generated;

Fig.12 is a diagram showing an example of the conversion of a hybridization-level pattern;

Fig.13 is a diagram showing an example of how a similarity pattern is displayed;

Fig.14 is a diagram showing another example of how a similarity pattern is displayed;

Fig.15 is a diagram showing another example of how a similarity pattern is displayed;

Fig.16 is a diagram showing yet another example of how a similarity pattern is displayed;

Fig.17 is a diagram showing an example of how a similarity pattern matrix is displayed; and

Fig.18 is a diagram showing a similarity pattern matrix shown together with the results of cluster analysis.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The preferred embodiments of the present invention will now be described with reference to the accompanying drawings. While the present invention is described by means of exemplary examples, in which probe biopolymers and sample biopolymers are both DNA, it will be appreciated that the present invention is not limited to hybridization experiments where DNA probes are hybridized to sample DNA but is also generally applicable to hybridization experiments which employ combinations of other biopolymers such as hybridization between DNA probes and sample DNA binding proteins, or hybridization between monoclonal antibodies and sample proteins.

Fig.1 shows a schematic diagram of a system in accordance with the present invention. The system includes a biochip data storage unit 100 for storing biochip data (name, definition information, experiment information or the like), data concerning levels of hybridization for each DNA probe on a biochip (levels of hybridization, spot images or the like), information concerning DNA probes (name, definition information, DNA sequences or the like) and other biochip data,

a display 101 for visualizing and displaying the stored data, a keyboard 102 for entering values into the system or performing selection, an input device 103 such as a mouse, and data processing unit 104 for processing data, the processing including calculating the similarity scores of DNA sequences, generating or sorting the similarity patterns.

Fig.2 is a flow chart schematically illustrating one example of a process for displaying a similarity pattern in accordance with the present invention. The process is described step-by-step with reference to the flow chart.

First, in step 200, a plurality of DNA probes to be investigated are designated (subject DNA probes). In step 201, information is obtained concerning each of the subject DNA probes.

This means that DNA probe data (e.g., names of the DNA probes, probe definition information, and DNA probe sequences) is retrieved from the biochip data storage unit 100 for each of the subject DNA probes selected in step 200. Next, in step 202, the similarity scores are calculated between two DNA probes of the subject DNA probes for all of possible combinations. The similarity scores are calculated between the same DNA probes, as well as between two different DNA probes.

Next, in step 203, biochips to be investigated are designated (subject biochips). A single or multiple subject biochip(s) may be selected. In step 204, data is obtained

concerning levels of hybridization for each subject DNA probe on each of the subject biochips. This means that data is obtained concerning levels of hybridization for a subject DNA probe on a subject biochip from the subject DNA probes selected in step 200 and the subject biochips selected in step 203.

Next, in step 205, a single DNA probe to be of concern in the investigation is selected (object DNA probe), and the subject DNA probes are sorted based on the similarity scores between the subject DNA probes and the object DNA probe. Alternatively, the process may proceed to step 206 in which data concerning levels of hybridization for a subject DNA probe on a subject biochip is processed by means of, for example, cluster analysis, and the subject DNA probes are clustered and sorted. Finally, in step 207, the results of the above processing are displayed for each of the subject DNA probes in the order obtained from the sorting in steps 205 or 206. If step 206 is taken, then additional information resulting from the data-processing, if any, is also displayed.

Processing performed in each of the steps shown in Fig.2 will now be described in detail by means of examples as shown in Figs.3 through 18.

Fig.3 schematically shows an example of data structure stored in a biochip data storage unit 100. Biochip IDs, biochip names, biochip definition information and biochip experiment information are stored in a biochip table 300. DNA probe IDs,

DNA probe names, DNA probe definition information and DNA probe sequences are stored in a DNA probe table 301. Also, biochip IDs, DNA probe IDs, levels of hybridization and spot images are stored in a hybridization-level table 302.

Figs.4, 5 and 6 show examples of biochip data stored in the biochip table 300, DNA probe data stored in the DNA probe table 301, and hybridization-level data stored in the hybridization-level table 302, respectively.

In step 200 in Fig.2, subject DNA probes are selected from the DNA probe data table 301 stored in the biochip data storage unit 100. As an example, a case in which DNA probes designated by DNA probe IDs 1 to 5 are selected from the DNA probe data table 301 to serve as the subject DNA probes is described. In step 201, DNA probe data (*e.g.*, DNA probe name, DNA probe definition information, and DNA probe sequences) are retrieved from the DNA probe data table 301 for each of the DNA probes selected in step 200 and designated by the DNA probe IDs 1 to 5.

Fig.7 shows an example of similarity scores calculated in step 202. Homology scores (*i.e.*, similarity scores) are determined according to the Smith-Waterman method using the DNA sequences of the DNA probes designated by the DNA probe IDs 1 to 5 that are retrieved in step 201. As shown, the similarity scores are displayed in a matrix-like form.

In step 203, subject biochips are selected from the

biochip data table 300 shown in Fig.4. For example, the biochip designated by the biochip ID 1 may be selected from the biochip data table 300 to serve as a subject biochip, or the biochips designated by the biochip IDs 1 to 3 may be selected to serve as subject biochips. In step 204, for example, data records are selected from the hybridization-level data 302a, 302b,..., based on DNA prove IDs of the subject DNA probes and biochip IDs of the subject biochips. Hybridization-level data (*e.g.*, hybridization level and spot images) are then retrieved for each record.

The similarity scores calculated in step 202 are displayed in square patterns (similarity patterns) in which, for example, different depths in a color are assigned to different values of the similarity score. As shown in Fig.8, a color gradient 603 is provided that corresponds to values ranging from a minimum value 602 to a maximum value 601 of the calculated similarity score. The color depths corresponding to similarity scores are determined and converted to a similarity pattern 604.

Specifically, as shown in Fig.9, similarity scores 605 obtained for DNA probes of DNA probe IDs:1 to 5 in DNA probe data table 301 with respect to the DNA probe of DNA probe ID:1 are converted to form a similarity pattern 606. The similarity patterns may be displayed as a similarity pattern matrix by arranging them in a matrix-like form as in the example of

similarity score calculation shown in Fig.7. Fig.10 shows an example of the similarity pattern matrix.

The hybridization levels obtained in step 204 are displayed in square patterns (hybridization-level patterns) in which, for example, different depths in a color are assigned to different values of the hybridization level. As shown in Fig.11, a color gradient 703 is provided that corresponds to values ranging from a minimum value 702 to a maximum value 701 of the hybridization level. The color depths corresponding to values of hybridization level are determined and converted to a hybridization level pattern 704.

For example, as shown in Fig.12, hybridization levels 705 obtained for DNA probe IDs: 1 to 5 in the hybridization-level data 302a shown in Fig.6 are converted to form a hybridization-level pattern 706. In this example, since the value range measurable by a typical biochip measurement device is in the order of two bytes, the minimum and maximum values 702 and 701 are set to 0 and 65535, respectively, so that the values can be expressed in the measurable range. However, it will be appreciated that, in practice, a hybridization pattern by which differences in values can be made visually more distinctive can be generated by employing the range of hybridization level to be processed.

Figs.13 to 18 show examples of how the results of the process can be displayed. In the following examples, data is

presented in a line for each of the subject DNA probes, making it easy to look at the information through the subject DNA probes. Data to be presented for each of the subject probes include DNA probe data 806, hybridization-level data 807 and a similarity pattern 808.

Fig.13 shows an example of a displayed data image in which Probe 3 in the DNA probe data table 301 in Fig.5 is used as an object DNA probe. In a displayed image 801, DNA probe data 806, which includes DNA probe names and DNA probe definition information, and hybridization-level data 807, which comprises spot images on a single subject biochip, are arranged adjacent to a similarity pattern 808, which has resulted from the conversion of similarity scores with respect to the object DNA probe. The name of the subject biochip for the hybridization-level data 807 is shown in a separate window 805. It can be seen from the displayed image 801 that Probe 5 has a higher similarity to Probe 3(objective probe) than do the other subject probes and that the hybridization level of Probe 5 is lower than that of Probe 1. This indicates that it is unlikely that the sample DNA corresponding to Probe 3 (objective probe) has been miss-hybridized to the other subjective probes including Probe 5.

Fig.14 is a variation of the displayed image shown in Fig.13. In a displayed image 802, DNA probe data 806, which includes DNA probe names and DNA probe definition information,

and hybridization-level data 807, which comprises a hybridization level pattern on a single subject biochip, are arranged adjacent to a similarity pattern 808, which has resulted from the conversion of similarity scores with respect to the object DNA probe. The name of the subject biochip for the hybridization-level data 807 is shown in a separate window 805.

Similarly, Fig.15 shows another example of displayed data image in which Probe 3 in the DNA probe data table 301 in Fig.5 is used as an object DNA probe. In a displayed image 803, DNA probe data 806, which includes DNA probe names and DNA probe definition information, and hybridization-level data 807, which comprises hybridization level patterns on a plurality of subject biochips, are arranged adjacent to a similarity pattern 808, which has resulted from the conversion of similarity scores with respect to the object DNA probe. The names of the subject biochips for the hybridization-level data 807 are shown in a separate window 805.

Fig.16 shows an example of a displayed data image in which the DNA probe with the DNA probe ID: 1(probe name: Probe 1) in the DNA probe data table 301 prepared in Fig.5 is used as an object DNA probe. The names of the subject biochips for the hybridization-level data 807 are shown in a separate window 805. A framed portion 809 in the hybridization-level data 807 indicates that the hybridization level of Probe 1 is very close

to that of Probe 2 in Chip 2. A framed portion 810 in the similarity pattern 808 indicates that the sequences of the DNA probes, Probe 1 and Probe 2, are very similar to one another. Together, these portions suggest the possibility of miss-hybridization in a readily recognizable manner.

Fig.17 shows an example of a displayed data image using a similarity pattern matrix. DNA probe data 806 including DNA probe names and DNA probe definition information, as well as hybridization-level data 807 comprising hybridization level patterns on a plurality of subject biochips, are displayed. In addition, a similarity pattern matrix 901 is arranged adjacent thereto in which similarity patterns are presented for all of the possible combinations between the subject DNA probes in a matrix-like manner. The names of the subject biochips for the hybridization-level data 807 are shown in a separate window 805. In this manner of displaying the image, the relationships can apparently be seen between the DNA probes with similar DNA sequences throughout the entire subject DNA probes, with respect to the hybridization-level data, for all of the DNA probes including the object DNA probe.

Fig.18 shows an example of a displayed data image in which a similarity pattern matrix is presented in combination with the results of the cluster analyses. In step 206 in Fig.2, DNA probes are clustered with respect to the changes in the hybridization level on each biochip, based on hybridization

level of the subject DNA probes on a subject biochip. Likewise, biochips are clustered with respect to the changes in the hybridization level for each DNA probe. The results are shown as tree diagrams 1001, 1002, respectively.

As shown, the subject DNA probes are sorted based on the results of the clustering with respect to the DNA probes. For each subject DNA probe, DNA probe data 806, which includes a DNA probe name and DNA probe definition information, and hybridization-level data 807, which comprises hybridization level patterns on a plurality of subject biochips, are displayed. In addition, a similarity pattern matrix 901 is arranged adjacent thereto in which similarity patterns are presented for all of the possible combinations between the subject DNA probes in a matrix-like manner. The names of the subject biochips for the hybridization-level data 807 are shown in a separate window 805.

In this manner of displaying the image, it is known for all of the subject biochips whether the results of the data analysis of the hybridization-level data reflect the physical properties of the DNA probe (*i.e.*, DNA sequence of the DNA probe) due to miss-hybridization or the biological properties of the sample (the manner in which the DNA sequences are present in the sample). In the example shown in Fig.18, it can be seen from the similarity pattern matrix 901 that Probe 1 and Probe 2 have DNA sequences that are very similar to one another while

the tree diagram 1001 indicates that the probes have rather different properties from one another (Probe 1 is more closely related to Probe 4). This suggests that the results of the analyses are reflecting the biological properties, rather than the physical properties, of the DNA probes.

As described above, the data images can be displayed in the manners shown in Figs.13 to 18 by following the flow of the processes shown in Fig.2. Also, after the results of the processing have been displayed, the object DNA probe may be replaced in step 208, and the data images are displayed in the same manner for the different object DNA probe. This makes it possible to verify the accuracy of hybridization.

Accordingly, the present invention provides a convenient display method which allows for the verification of the accuracy of hybridization experiments in the art of biochips by making use of DNA sequences of the subject DNA probes.

While there has been described what are at present considered to be preferred embodiments of the present invention, it will be understood that various modifications may be made thereto, and it is intended that the appended claims cover all such modifications as fall within the true spirit and scope of the invention.